Архитектура на системна мрежа за колективна комуникация ГРАНД-КЛОС в суперкомпютри

Десислава Иванова

As innovative approaches and last trends engineering computing methodologies the demands for high-speed computational resources are growing at a rapid rate. The parallel computer performance of system network architecture of collective communications is a critical factor influencing the speed up and performance of supercomputers which design the next generation architecture of modern supercomputers. In this paper modern collective network architecture of "Grand Clos" topology design is proposed to cover the needs of efficient and high-speed communication on supercomputers. The communication performance of the proposed new architectural design for collective communications is evaluated on the basis of parallel simulation models using the framework OMNET++ (C++, MPI). The simulation tests are running on IBM HS22 Blade Center, located at High-Performance and GRID Computing Laboratory, Computer Systems Department, Technical University of Sofia. Analysis of simulation results has been performed and presented in the paper.

Key words: *Grad Clos Architectural Design, Collective Networks, Supercomputers, OMNeT++, MPI, Communication Performance Evaluation*

1. ВЪВЕДЕНИЕ

Голяма част от съвременните компютърни приложения изискват все по-високи скорости на информационната обработка, надхвърлящи възможностите на последователните компютри. През последните години се наблюдава множество от интензивни научни изследвания и разработки в областта на компютърните системи и софтуерното инженерство свързани с повишаване на компютърната производителност. Суперкомпютрите свързват голям брой процесорни възли, под формата на многостъпални хибридни мрежови архитектури. Едно от предизвикателствата при работа със суперкомпютри е необходимостта от обмен на информацията между процесорите, с цел получаване на крайно решение от изчислението на терамащабни задачи. За осъществяването на тази цел се използват високоскоростни колективни комуникационни мрежи [1, 2].

Архитектурните аспекти и характеристики на колективните мрежи в суперкомпютрите до голяма степен определят пиковата производителност на високопроизводителните компютърни системи. Колективната мрежа представлява метод за постигане на висока скорост, ниска латентност и възможност за осъществяване на глобални колективни комуникации. В колективните мрежи на суперкомпютрите са включени маршрутизиращи устройства, които се свързват с възлите в мрежата, за да улеснят преноса на данни и изпълнението на глобални операции [3, 4].

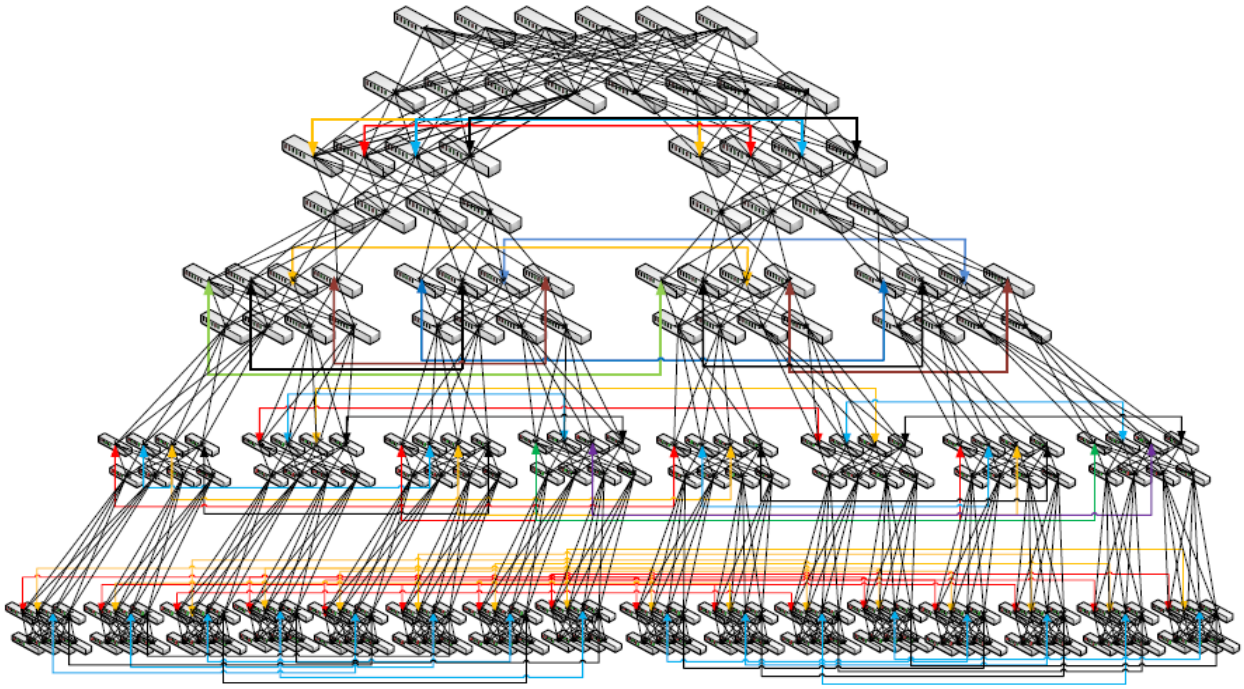
В настоящата статия се предлага нов иновативен дизайн на системна мрежа за колективна комуникация с топология ГРАНД-КЛОС за суперкомпютри. Оценка на комуникационната производителност на предложеният дизайн на колективна мрежа за суперкомпютри е извършена на база компютърни симулации, проведени на високопроизводителната платформа IBM HS22 Blade Center, локализиращ се в лабораторията по „Високопроизводителни компютърни системи и ГРИД технологии", към катедра „Компютърни Системи", Технически Университет – София.

2. АРХИТЕКТУРА НА СИСТЕМНА МРЕЖА ЗА КОЛЕКТИВНА КОМУНИКАЦИЯ ГРАНД-КЛОС

Изследванията и усилията са насочени към проектиране и изграждане на нов иновативен архитектурен дизайн за колективна комуникация с топология ГРАНД-КЛОС за суперкомпютри, който да осигурява добри стойности на комуникационните параметри. Предложеният нов архитектурен дизайн за колективна комуникация с

топология ГРАНД-КЛОС е хибриден дизайн, изграден от мрежи с топологии „Дебело Дърво“ и „Клос“. Като хибриден модел, новата ГРАНД-КЛОС мрежа за колективна комуникация има значителни предимства, в сравнение с класическите „Дебело Дърво“ и „Клос“ мрежови топологии, Фигура 1.

Новият дизайн на топология се състои от няколко свързани Клоса, в топология „Дебело Дърво“. Най-важната част на новият архитектурен дизайн е наличието на хоризонтални връзки между възлите на „Дебело Дърво“ на същото ниво. Тези хоризонтални връзки минимизират комуникационните разходи и необходимия брой стъпки при колективните комуникации. Също така, тези хоризонтални връзки намаляват натоварването на „корена“ (root node). Обикновено, чрез прилагане на бисекция, мрежата „Дебело Дърво“ се разделя на две части, при което целият трафик от едната част минава през корена, за да достигне възлите от другата част.



Фигура 1. Архитектура на системна мрежа за
колективна комуникация ГРАНД-КЛОС

В предложения нов архитектурен дизайн „Гранд-Клос“ за колективна комуникация, хоризонталните връзки помагат в колективните комуникацията между отделните части на мрежата, намаляват натовареността на корена на мрежата и броя на стъпките при колективните комуникации, което води до увеличаване на пропускателната способност и до намаляване на латентността в ГРАНД-КЛОС мрежата. В допълнение, връзката между Клосовете е надеждна, защото всеки има връзка с другите посредством кросбар. Основната идея е да се използват предимствата на двете топологии в изграждането на новия хибриден дизайн за колективна комуникация с допълнителна иновация свързана с добавяне на хоризонтални връзки между възлите на „Дебело Дърво“ на същото ниво.

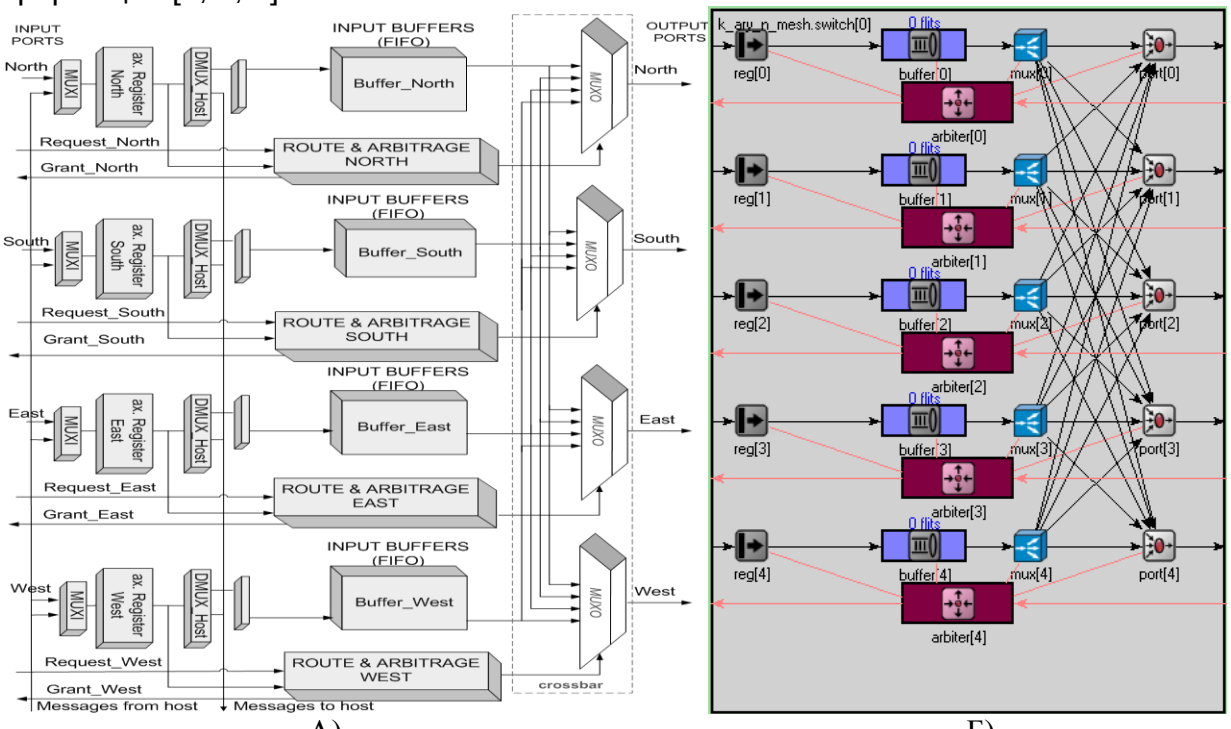
Един Клос се състои от 8 възли (2 реда, 4 възли). Коренът (Клос, върха на дървото) 6 възли на първия ред и 8 възли на първия ред. Мотив е нуждата от повече низходящи връзки. Той разполага с 16 връзки към второ ниво (8 за лявия клон и 8 за дясната част). Всяки Клос разполага с 8 връзки към възходящото ниво, 16 връзки към низходящото ниво и 8 хоризонтални връзки. Така, общият брой на възли е 254 (30 малки клосове, изградени от 8 възли и 1-големи Клос (Root Clos) изграден от 14 възли).

Друг важен аспект на предложения архитектурен дизайн са хоризонталните връзки между клосовете на същото ниво. За всеки клос, възли 2 и 4, (сини и черни стрелки), както и възли 1 и 3 (червени и оранжеви стрелки) са свързани към различни клосове от двете части на „Дебелото Дърво“. Това помага да се намали броят на стъпките на комуникация (hops), по време на предаване на съобщенията.

Тази хибридна структура позволява да бъдат избегнати недостатъците на двете отделни мрежови топологии, както и да се съчетаят техните предимства. С тази хибридна топология се увеличава надеждността на мрежата – „коренът” вече не е слабо място; също така се избягва ситуацията, в която при повреда на възел, неговите наследници губят връзка с мрежата – комбинацията от двете топологии осигурява връзка на възлите от по-долно ниво с останалата част от мрежата, дори при повреда на възела „родител”. Едновременно с това тази топология ни позволява лесно да откриваме повреди в съобщителната среда, както и да се възползваме от лесното осъществяване на колективната комуникация. Едновременно с това, комбинацията от двете топологии осигурява алтернативни маршрути при предаването на пакетите и намалява вероятността от възникване на блокиране.

3. АРХИТЕКТУРА НА КОМУТАТОР ЗА КОЛЕКТИВНА КОМУНИКАЦИЯ

Архитектурата на комутатора за колективна комуникация е изграден от три стъпала: входни регистри, FIFO опашки (буфери) и неблокиращ кросбар, имплементиран като напълно свързани мултиплексори и демултиплексори. Входния мултиплексор, който избира посоката на пренасочване на трафика – към хоста или към някой от съседните комутатори (към хоста с по-нисък приоритет). Входните регистри (reg(*)) записват един флит и извличат неговата маршрутизираща информация [6, 7, 8].



А) Б)

Фигура 2. А) Архитектура на комутатор за колективна комуникация
Б) Симулационен изглед на комутатор за колективна комуникация

Отделните основни (прости) модули, както и инстанциите на комутатора в мрежата, са свързани посредством предефинирани връзки - канали (channels). Графично изображение на описания съставен модул, изграждащ комутатора, Фигура 2.

4. МРЕЖОВА ПЛАТФОРМА ЗА СИМУЛАЦИИ OMNeT++

Една от най-важните задачи при проектирането на компютърните мрежи е оценката на техните комуникационни параметри за производителност и динамично поведение. Моделирането и симулацията на мрежи за последно поколение суперкомпютри е сложна и комплексна задача, която изисква проектиране на прецизен симулационен модел и тестване на модела при различни сценарии [5].

OMNeT++ по същество представлява набор от софтуерни инструменти и библиотеки, подпомагащи разработката на симулационни модели за компютърни мрежи и протоколи, но практически може да се използва за изготвяне на всякакви модели. OMNeT++ е симулационна среда, която включва специфични библиотеки (simulation framework and library). Изградена е от отделни компоненти, наречени модули. Главното и предназначение е изграждането на мрежови симулации от най-различен вид. Под „мрежови“ тук се има предвид широкия смисъл на думата, който включва жични и безжични комуникационни мрежи, вградени мрежи върху чип (NoCs) и други. OMNeT++ включва Eclipse базирана графична среда за разработка и някои други допълнителни инструменти за улеснение работата на разработчиците. Съществуват и разширения за симулиране в реално време, емулиране на мрежи, както и възможност за използване на допълнителни програмни езици като Java и C# и възможност за интегриране на бази данни, както и много други допълнителни функции.

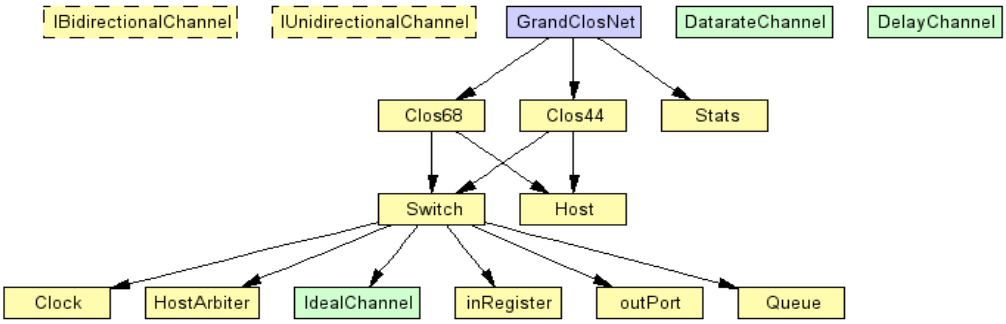
5. ПАРАЛЕЛНИ МОДЕЛИ И ИМПЛЕМЕНТАЦИЯ В OMNeT++

5.1. Експериментална платформа

За проектиране и провеждане на симулационните тестове се използва високопроизводителната платформа IBM Blade Center, базирана на Blade Server HS22, в лабораторията по „Високопроизводителни Компютърни Системи и ГРИД Технологии“ към катедра „Компютърни Системи“ на Технически Университет – София.

5.2. OMNeT++ модели и диаграми на ГРАНД-КЛОС архитектурата за колективна комуникация

Представените в този раздел на статията OMNeT++ диаграми показват отношенията между простите и съставните модули в архитектурата на системната мрежа за колективна комуникация ГРАНД-КЛОС за суперкомпютри, модулните интерфейси, мрежите, връзките между модулите и техните интерфейси.



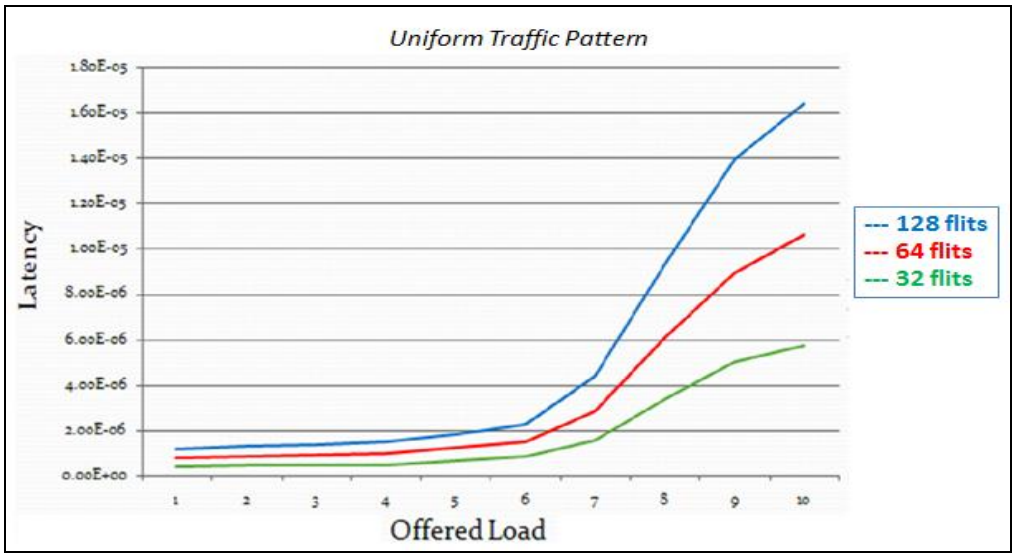
Фигура 3. Йерархия на модулите в OMNeT++ за системната мрежа за колективна комуникация ГРАНД-КЛОС

За всеки Клос, броят на хостовете е равен на броя на комутаторите, така че един хост е свързан към всеки комутатор. Фиг. 3 показва компоненти на комутатора. Модулът часовник обработва „Cmessages“. „Cmessages“ представляват всички

събития и съобщения в симулацията. Модулът HostArbiter управлява комуникацията между комутатора и хостовете свързани с него. Всички останали компоненти на комутатора са свързани с архитектура, описана в точка 3.

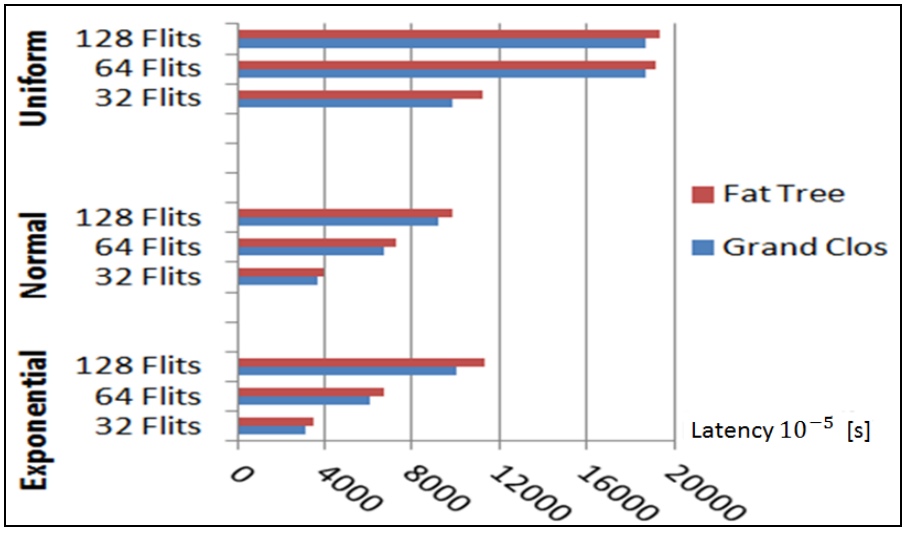
6. СИМУЛАЦИОННИ РЕЗУЛТАТИ

Симулационните тестове проведени за архитектурата на системната мрежа за колективни комуникации ГРАНД-КЛОС са извършени за три различни размера на предаваните в мрежата пакети: 32, 64 и 128 флита и три различни разпределения на трафика в мрежата равномерно, нормално (или гаусово) и експоненциално. За да се тества всяка комбинация от различните размерите на пакети и разпределенията на трафика в мрежата се извършени 9 симулации. Всяка симулация продължава около 2 часа на хардуерната платформа IBM HS22 Blade Center за 20000 предавани пакети. Следващата графика показва средната латентност при предаването на пакети с различен размер при равномерно разпределение на трафика.



Фигура 4. Латентност при равномерно разпределение
на трафика в мрежата и различен размер на пакетите

Видими са разликите в получените стойности на латентността. От получените резултати се наблюдава фактът, че за симулациите проведени за размер на пакетите 32 флита, стойностите на латентността за всички разпределения на трафиците в мрежата са най-ниски.



Фигура 5. Симулационни резултати за мрежи с

топологии ГРАНД-КЛОС и Дебело Дърво

Симулационните резултати визуализирани на Фиг.5 показват, че предложеният нов архитектурен дизайн на мрежа за колективна комуникация с топология ГРАНД-КЛОС за суперкомпютри постига 15% по-ниска латентност за трите изследвани разпределения на трафика в мрежата и трите размерности на пакетите. Постигнатата комуникационната производителност обяснява използването на допълнителните разходи за хоризонталните връзки в предложения нов архитектурен дизайн на системна мрежа за колективна комуникация с топология ГРАНД-КЛОС.

ЗАКЛЮЧЕНИЕ

В тази статия се предлага нов иновативен архитектурен дизайн на системната мрежа за колективна комуникация ГРАНД-КЛОС. Разработени са паралелните модели и са проведени паралелни симулации за три различни размера в пакетите: 32, 64 и128 флита и три различни разпределения на трафика в мрежата: равномерно, нормално (Гаусово) и експоненциално. Всички експерименти са проведени на IBM Blade Center, в лабораторията по „Високопроизводителни Компютърни Системи и GRID Технологии, към катедра „Компютърни Системи“, Техническия университет - София.

В бъдеще, изследванията ще бъдат ориентирани към провеждане на симулации с цел да се сравни иновативна ГРАНД-КЛОС архитектура за колективна комуникация и имплементираната в суперкомпютъра IBM BlueGene/P колективна мрежа, която е с топология „Дебело Дърво – 3D Тороид“. Допълнителните тестове ще дадат възможност да се направи пълна оценка на комуникационната производителност и да се сравни и оцени ефективността на предложения нов архитектурен дизайн ГРАНД-КЛОС.

ПРИЗНАТЕЛНОСТ

Статията е публикувана с финансовата подкрепа на проект №132ПД0046-09.

ЛИТЕРАТУРА

[1] Ahmad Faraj, Sameer Kumar, Brian Smith, Amith Mamidala, John Gunnels, Philip Heidelberger: MPI Collective Communications on the Blue Gene/P Supercomputer-Algorithms and Optimizations, IBM T. J. Watson Research Center.

[2] James Milano Gary L. Mullen-Schultz, Gary Lakner: BlueGene-red book: Blue Gene/L: Hardware Overview and Planning.

[3] P. Borovska. Computer systems. Sofia; Bulgaria: Ciela, ISBN 954-649-633-2 (in Bulgarian), 2009.

[4] Duato, J., Yalamanchili, S., Lionel M. Interconnection networks: An engineering approach, Morgan Kaufmann Publishers, ISBN 1-55860-852-4, 2002.

[5] http://omnetpp.org/doc

[6] Pl. Borovska, O. Nakov , D. Ivanova, K. Ivanov, G. Georgiev: Communication Performance Evaluation and Analysis of a Mesh System Area Network for High Performance Computers. 12-th WSEAS International Conference on Mathematical Methods, Computational Techniques and Intelligence Systems (MAMECTIS’10), Kantaoui, Sousse, Tunisia, May 3-6, 2010, ISBN: 978-960-474-188-5, pp. 217-222.

[7] P. Borovska, O. Nakov, D. Ivanova, A. Ruzhekov, Halil Mohamed: A Comparative Analysis of Next Generation High-End Switch Architectures. Fifth International Conference "Computer Science", Bulgaria, Proceeding, pp. 7-12, 5-6 November 2009

[8] P. Borovska, D. Ivanova, K. Ivanov, G. Georgiev: Generalized Simulation Model of a Switch for High-Speed Interconnection Networks, Sixth International Scientific Conference Computer Science’2011, Ohrid, Macedonia, pp. 17-22, 01 - 03 September 2011

За контакти:

Асистент Десислава Иванова, Катедра “Компютърни системи“, Технически Университет - София, тел.: +359 2 965 22 24, e-mail: d_ivanova@tu-sofia.bg

Докладът е рецензиран.